The Making Of the IGARSS '94 Proceedings CD-ROM

by
Mike Martin, Ann Bernath/Jet Propulsion Laboratory
Mark Takacs/Sterling Software

The International Geoscience and Remote Sensing Symposium (IGARSS) '94 was held at Caltech, in Pasadena, CA, August 8-12. The conference was attended by about 800 scientists and the agenda included nearly one-hundred sessions with ten presentations per session. For the first time, the conference proceedings were provided on a CD-ROM disc instead of the traditional hard-copy format. Attendees had to specifically request (and pay $65) to purchase the four-volume, 15 lb. printed proceedings (440 attendees did so).

The IGARSS CD challenge was to provide equivalent content to the printed proceedings on a CD-ROM disk. Anyone familiar with electronic publishing will understand that this is a significant and possibly impossible challenge. The display size of a computer screen may vary from 640 pixels (wide) by 480 pixels (high) to 1024 by 768. A printed page needs to be scanned at about 300 dots per inch to provide an adequate rendition which results in an image size (for an 8-1/2 x 11 inch page) of 2550 by 3300 pixels. Thus a standard computer display can view only a pathetically blurry version of an entire page, or a full-resolution version of a tiny part of a page. With this realization in mind, the requirements for the IGARSS CD-ROM were limited to the following:

- provide all abstracts in full-text form with some text search capability to assist in locating papers of interest

- provide monochrome scanned versions of all papers and a mechanism to print on demand the scanned papers

- include author submissions of gray scale or color images

The final architecture of the CD-ROM exceeds most of the objectives:

- provides the full text of all abstracts via Mosaic, but not with full-text search on the CD-ROM (there just wasn't time to package standalone WAIS on the CD-ROM); rather specialized indexes were provided by author, paper number, topic, neat papers, etc.

- provides papers in Adobe Acrobat Portable Document Format (PDF) in one of several forms accessible through the Mosaic interface (one paper at a time) or in a 3550 page PDF file, which includes the table of contents with hyperlinks to the papers and bookmarks which outline the conference agenda. The formats of the papers may be any one of the following:

    - rich documents just as the author prepared them (many of these are truly magnificent examples of what electronic publishing can be)

    - stripped down versions of the above, but missing the graphics and images, which may be available through the Mosaic interface.

    - monochrome scanned TIFF images of each paper, in acceptable, but not great quality due to our paper scanning procedures.

  - strange-looking renditions of the papers interpreted by Image Alchemy from postscript files and presented in image format, rather than text.

- author submissions of gray scale of color images for viewing via Mosaic helpers.

Figure 1 IGARSS source material, papers, floppies and abstracts

Figure 2 IGARSS Proceedings and IGARSS'94 CD-ROM

The IGARSS challenge was exacerbated by the desire to support three computer platforms (Macintosh, Windows and Unix workstations) with the same, or at least a similar, interface. Most conference proceedings to date rely on delivery software that works on only MS-DOS or Windows-based platforms. The science community, however, is pretty evenly divided across the selected platforms. The delivery mechanisms chosen (Mosaic and Adobe Acrobat) work consistently across the target platforms.

Document Types

It is helpful to summarize the various formats used for document preparation and delivery. All of the following formats were encountered in the IGARSS CD production.

Original document - A file or files prepared in some word processor format, unique to a particular program and computer system. May have dependencies (special fonts or “style sheets” for formatting) which impair portability. May have embedded graphics and images or just place holders for external image files which are provided to a publisher in hard-copy format. Requires word processor software to view.

Postscript document - A postscript printer file produced by “printing” the original document to a file instead of a printer. Most computers provide this capability in the printer options dialog. This format should retain the information content of the original document when printed, but may not retain the full-text or provide access to embedded graphics or images. Some word processors and utility programs can display postscript files on the screen. Postscript files are encoded ASCII text and are normally fairly large (several hundred kilobytes per page).

Scanned document - Images of pages of the document, almost universally stored in Tagged Image File Format (TIFF) format. A resolution of 300 dots per inch is commonly used for document scanning. There are substantial incompatibilities between TIFF formats and display software especially across platforms. Scanning can be monochrome (one-bit per pixel, good for text, figures and equations); grayscale (8-bits per pixel, good for grayscale images or shaded figures) or color (24-bits per pixel, good for color graphics only). Note that there is no optimal scan for mixed pages and to obtain an optimal representation of a page with text, gray scale and color three separate scans are required! Uncompressed pages yield 1 megabyte monochrome, 8 megabytes gray scale or 24 megabytes color. Compressed page sizes are on the order of 30 kilobytes per page for monochrome and 150 kilobytes per page for gray scale, several hundreds of kilobytes for color.

HTML document - ASCII text marked-up with simple SGML/HTML tags to indicate formatting and hyperlinks to anchor points within a document or to external files which can contain text, images, and other data types. This format retains the flavor of the original document but is not an exact duplicate. It is easy to automate the production, formatting and linking of HTML documents. HTML has already produced the most explosive growth in information distribution in history and could well serve as the uniting force in the convergence of printed and electronic publishing.

PDF document - Encoded ASCII text file containing Portable Document Format (PDF) mark-up and highly compressed graphics and images. This format retains the information content of the original document. The resulting files are several times smaller than the source postscript file. Produced by “distilling” a postscript file using the Adobe Acrobat Distiller program or “printing” to a PDFWriter printer driver. The document creator can use a special program (Acrobat Exchange) to build hypertext links (by hand) or mark-up the PDF with PDFMark commands.

Abstract processing.

Optical Character Recognition (OCR) was performed on approximately 1100 abstracts using a Kurzweil 5200 OCR system. About 50 abstracts were OCR’d at a time, with each batch taking about 30 minutes. Ninety percent of the resulting text files contained only minor errors and could be hand-edited in about 3 minutes per abstract. The major difficulties in abstract processing were the formatting of author names and addresses and the use of embedded mathematical symbols (despite requesting that authors not put them in abstracts!).

Bad abstracts.

About 120 abstracts contained so many errors that the editors felt it was easier to retype the whole page (7 minutes per abstract). Instead of this massive retyping, we decided to try an experiment. The abstracts were re-scanned on Kodak Scanner/Microimager 990, which provides 120 page-per-minute throughput and can scan both sides of a document simultaneously. The scanning took just a few minutes. The scanned pages were then OCR’d with two different OCR programs (WordScan and OmniPage). The resulting text files were compared using a Macintosh program called DocuComp II. This program merges all the identical text and highlights the differences. Incredibly, the combined success rate of the two OCR programs was very close to 100 percent. It then took only a few minutes per abstract to discard incorrect text. This technique was summarized in the article “Improving Optical Character

Recognition Performance” in the April, 1994 issue of the Information Systems Newsletter.

Hardcopy paper processing.

The papers were received in random order in various sizes, but mostly 8-1/2 x 11 pages. The first step was sorting the 720 papers into some order prior to scanning. The scanning was done on the Kodak Scanner/Microimager 990 scanner that was used to scan the bad abstracts. The machine setup had to be modified for our scanning which took about an hour. The scanning process itself was similar in complexity and speed to inserting a stack of pages into a photocopier and pushing the "start" button. However, the secondary processing to produce our TIFF output files took more on the order of 20 seconds per page (3 pages per minute). This was substantially slower than the abstract conversion had been, due to the greater information density of the pages. The resulting TIFF files were about 38 kilobytes per page.

The greatest problem encountered in the scanning process was difficulty in spotting scanner misfeeds. Due to the sheer speed at which the scanner pulled the pages through, it was only possible to spot the grossest misfeeds. It was not possible to determine when the scanner had pulled two pages through as one page, which happened about two percent of the time.

Scanned paper processing.

The scanning resulted in 2250 TIFF images which had to be converted to 720 PDF files. This process consisted of three parts:

- Associating individual TIFF image files with a particular paper, a tedious manual process of opening at least the first page image and comparing it to the folder of papers that was scanned then recording the starting page number of the paper.

- Converting the TIFF images to PDF format and renaming the resulting files with the paper number, which was done automatically from scripts using the Image Alchemy program on a Sun workstation.

- Concatenating the individual PDF files into a single PDF file with Acrobat Exchange. This was done manually, one paper at a time, adding a page then saving, adding another page and saving, etc.

Nearly a third of the TIFF files scanned on the Kodak scanner were not readable when processed on the Sun workstation. These files were rescanned on the Kurzweil 5200 at 400 dpi. This produced images which were too large for Acrobat (which is limited to 2048 pixels). These files were converted to 300 dpi with image alchemy. The first time these PDF files were scrutinized was when the first review CD was made. Unfortunately, the conversion from 400 dpi to 300 dpi had rendered many of the pages unreadable. The conversion had to be redone, using optimization options which produced acceptable 300 dpi images.

Electronic Submissions.

Electronic submissions were received by anonymous FTP, e-mail, and on floppy disks. A special ftp site and a special IGARSS e-mail account were set up for the incoming IGARSS submissions. Approximately 153 papers were submitted electronically via FTP and e-mail. Approximately 125 more were submitted on floppy disks (70 percent PC, 30 percent Mac).

Electronic submissions were identified by paper number (a sequential number assigned to abstracts as they were received), author name, title and session id (a concatenation of day, time, room and sequence number). An explicit scheme for labelling submissions should have been specified in the author instructions. Several hours of unnecessary detective work were required to track down the proper paper numbers. The files received for each paper were placed in directories named by the paper number. Where available the e-mail addresses of the authors were kept in case questions arose later and to request a final review of the CD-ROM contents.

The e-mail and ftp directories needed to be watched over carefully. Unfortunately, during a business trip the e-mail directory filled-up the file system and was deleted in the ensuing recovery process. The nightly backups that were being performed turned out to be capturing only a static part of the file system not the e-mail directory. It was very embarrassing to have to request that the authors resend electronic submissions submitted for that period of time.

Authors had been asked to submit postscript files and/or ASCII text. Many also submitted word processor files, which turned out to be beneficial when some postscript submissions could not be processed. They were also allowed to submit up to one megabyte of images or other data files in any format they desired. These supporting files were placed in the directory with their associated paper, and in some cases the image files were converted to a standard format such as GIF. Some authors sent TeX and LaTeX files. One author sent a GML file which was easily converted to HTML (HyperText Markup Language). In fact, we were able to convert 18 papers to HTML. These papers were totally integrated with the rest of the Mosaic interface and did not require launching the Acrobat Reader.

Postscript File Processing.

Each postscript file had to be run through the Adobe Acrobat Distiller program to produce a PDF file. There is a way to distill multiple postscript files into one PDF which might have been useful had we known about it early in our processing. Of the 283 papers received in postscript format, 122 distilled correctly, 63 produced fatal errors, 60 resulted in papers that displayed so slowly as to be unusable, 26 were clipped (only a portion of the page was displayed) and a dozen or so were otherwise damaged in bizarre ways. Through a variety of tedious methods about 48 of the bad papers were rescued.

There was no discernible pattern to the fatal distiller errors. Most were *OffendingCommand* errors, but the commands which offended were different from paper to paper (AldusDict2, get, awidthshow, EJ, findfont, nostringval, dlenvelope, etc.). It was frustrating that a lot of the errors were close to or at the end of a paper, but the Distiller deleted any good pages when it flushed the job. Of those that did distill, some displayed very, very slowly. We traced this to the use of a TeX DVI (Device Independent output file) to Postscript utility that changed all the text to bit-map graphics. These files were replaced with the TIFF scanned versions or converted using Image Alchemy PS. The cropped pages were due to a divergence in the requirements for submissions for the printed proceedings and the electronic version. Printed copies were requested on 9-inch by 14-inch mark-up sheets. While most authors printed 8 1/2 x 11 pages then cut and pasted them onto the sheets, some produced their printed pages at 9 x 14 and then made postscript files of that size.

For some of the bad postscript files we also had the word processor files used to generate the postscript. In these cases we opened the files with Microsoft Word then printed to the PDFWriter printer driver. This technique allowed us to salvage a handful of the bad PDF files.

The rest of the bad postscript files were processed with Image Alchemy PS, (one page at a time) to produce PDF pages which were then concatenated into PDF papers. A good portion of these had to be re-scaled during the conversion so they would load in Adobe Acrobat Exchange without an out-of-bounds error. In this recovery effort, only the black and white conversion was successful. Trying to convert Postscript files containing gray-scale or color resulted in incredibly large files per page that were essentially unusable. Unfortunately, as a result, potentially beautifully rendered Postscript papers were reduced to black and white because of Adobe Acrobat's inability to process them through their Distiller product, and Image Alchemy PS's inability to produce a usable color file.

Mosaic Interface.

The IGARSS home page presents a variety of information about using the CD-ROM, about the conference and a set of menu access paths to get to the papers. The abstract for each paper serves as that paper's "home page," a stable central access point for all the information submitted with that paper. The abstract/home pages were automatically generated via a perl script which made revisions and style changes trivial. The script examined each paper's directory and constructed links to all files submitted with that paper. These often included the full-text PDF files, ASCII text-versions, code-fragments, and images or figures in GIF, JPG, or EPS format.

When the user selects a paper from one of the access paths the abstract text is displayed with an iconic link to the paper (stored in Acrobat format) and a menu of files submitted by the author. Selecting the paper link launches Adobe Acrobat and displays the paper. The Acrobat display ranges from excellent for papers that were formatted for a computer screen to miserable for many of the scanned papers. The key value Acrobat provides is the ability to preview then print papers of interest. The printed copy is a faithful (if not pretty) rendition of the contents of the original proceedings.

Selecting any of the associated image files in a display helper being launched (JPEGVIEW on the Mac, LVIEW in Windows and XV on a Sun) to display the image.

Review Process.

The best review process for the IGARSS contents was having the authors log onto the on-line www IGARSS home page and view their material. In early July we distributed copies of a draft CD to about 10 scientists at JPL. Unfortunately, most of the review group were also conference committee chairpersons which meant that they were already overloaded with work. The review pointed out only a handful of problems with the data, but forcefully demonstrated the inadequacy of our user documentation. Only one reviewer was able to get all the elements of the Mosaic interface installed.

Creating IGARSS.PDF.

The review process led us to believe that we should include a single Acrobat PDF file with the table of contents and all papers so that unsophisticated users could simply install Adobe Acrobat and open a single file to access the entire proceedings. This turned into a nightmarish task, requiring on the order of two full person-weeks of mindless drudgery to combine over 720 PDF files, find the page location of every paper and build (by hand) the links between the table of contents and the papers. The concatenation of PDF's was required because Acrobat 1.0 did not allow links to external files, only links within a file. Had we been using Acrobat 2.0 we would not have needed to perform the concatenation. The final IGARSS94 PDF file weighed-in at a hefty 233 Megabytes and 3550 separate pages, including a 60 page table of contents. We were concerned that the large Acrobat file would be too unwieldy to use. The first few papers accessed were slow to load and display. However, after the initial load the access and display speed increased noticeably. During the demonstrations using the igarss.pdf file, there was no noticeable difference between jumping to a paper on page 50 versus one on page 3300.

Customer Response.

During the conference, both the Adobe Acrobat interface and NCSA's Mosaic interface were available to users to browse the IGARSS '94 CD-ROM. A number of conferees used the CD to determine their agenda for the day. The majority of users preferred the Mosaic interface because the information on the screen was easier to read and the access paths were more obvious. Traversing the Acrobat document was not intuitive and the users could became lost in a sea of concatenated pages. It should be noted that Acrobat could support building access paths like those found in Mosaic, but only with a great deal of manual hyperlinking.

A major problem reported by CD-users was the fact that only one Unix platform, SunOS, could use the Acrobat Reader. HPs, SGIs, and other hosts could navigate within Mosaic, but they could not view the papers. Some Solaris users also had problems, even though we had successfully used the SunOS binary on a Solaris machine.

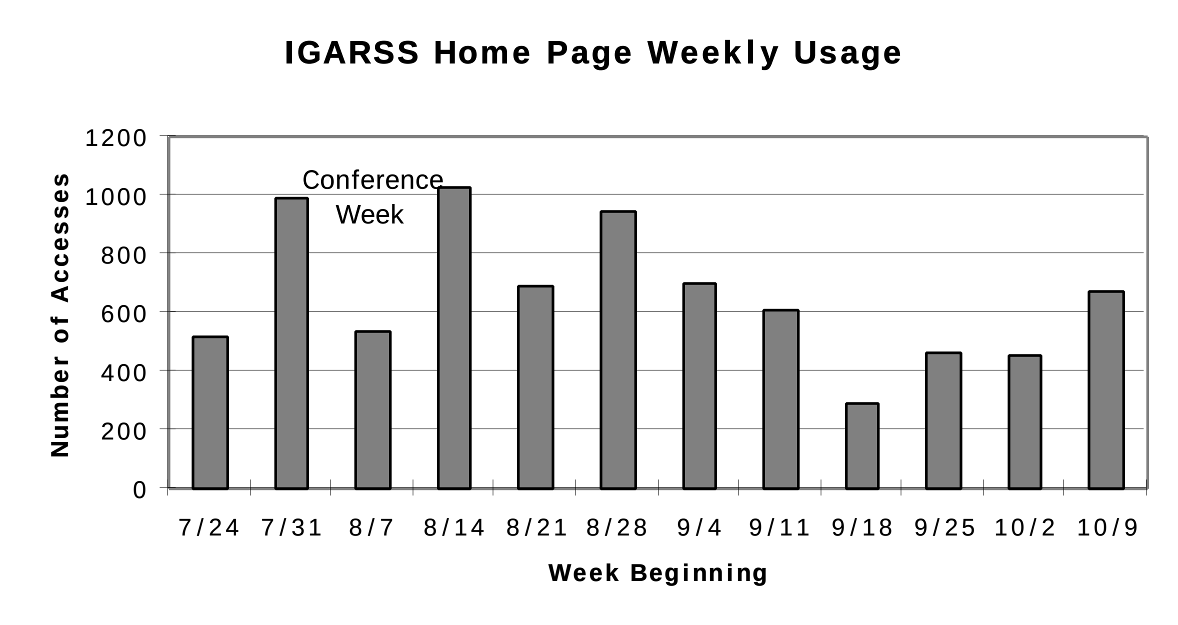

Figure 3 IGARSS Home Page Usage Statistics

The on-line WWW server was used regularly before, during, and after the conference and continues to be used as of this printing. There were spikes before and after the conference as people explored what they might anticipate at the conference and when they returned home and consulted the on-line version.

For the peak usage week, accesses were split between .gov, .com and European sites, with European and gov sites registering high consistently through all the weeks. Updated statistics are available at *http://stardust.jpl.nasa.gov/igusage*.

Resources.

The planned IGARSS CD-ROM budget was $12K for labor and $12K for CD production (2 CDs were initially planned, but only one was produced). The actual cost was $12K for labor, $2K for a 2 gigabyte hard disk (which proved invaluable) and $5K for the 1700 copies of the CD, including the color 12-page brochure. Digital Audio Development Corp. (DADC) did the brochure, mastering and replication and were nice enough to provide a 2-day turn-around at the 7-day turn-around price.

About another $8K of labor was provided by the Data Distribution Laboratory staff or voluntarily contributed on weekends and evenings in order to produce a great product. A rough estimate of three person-months is probably close to the total effort. The following table attempts to summarize the parts of the task by person hours.

Scan 1,100 abstracts	24 hours
Edit 1,100 abstracts	48 hours
Organize 720 papers (2250 pages)	16 hours
Scan 2250 pages/rescan 669 pages to TIFF format	24 hours
Process 126 floppy disk submissions	8 hours
Process 153 electronic submissions	16 hours
Distill postscript files to PDF	16 hours
Process TIFF files to PDF	16 hours
Produce HTML files from ASCII submissions	8 hours
Rescue bad submissions	40 hours
Produce documentation	24 hours
Produce Access Paths for Mosaic version	8 hours
Build scripts for abstract handling, etc.	16 hours
Produce individual PDF files	40 hours
Produce monster file IGARSS.PDF	40 hours
Inventory IGARSS.PDF file	24 hours
Build IGARSS.PDF hyperlinks	16 hours
Validate IGARSS.PDF	8 hours
Conduct review	16 hours
Database Editing and Maintenance	16 hours
Task Management (meetings, planning)	16 hours
Total	440 hours

The software used in the production on the IGARSS conference proceedings included: Adobe Acrobat Reader, Exchange and Distiller 1.0; NCSA Mosaic; WAIS; Image Alchemy (Sun version); UserLand Frontier (Mac); QuicKeys (Mac); FoxPro (Mac); Perl (Sun); Word (Mac); and MacLink Plus (Mac).

Conclusions.

The IGARSS CD-ROM team feels that the IGARSS CD-ROM provides at least a glimpse of what electronic conference proceedings can be. It's not everything we wanted but exceeds anything we've seen before in content and scale. Over seven-hundred abstracts and papers are available in a form equivalent to the printed proceedings (scanned pages). Several hundred papers are available in full electronic form, including many color submissions that are not found in the printed proceedings. The resulting CD-ROM is literally a treasure of readily accessible knowledge with supporting digital data sets that can be applied to almost any area of remote sensing or science education.

Copies of the IGARSS CD are available from:

The Institute of Electrical and Electronics Engineers, Inc.
445 Hoes Lane
P.O. Box 1331
Piscataway, NJ 08855-1331

Suggestions for future IGARSS CDs.

These are some of the things we would do differently if we were to produce another IGARSS CD-ROM.

- Require electronic submissions of abstracts with tags for automated processing. Mandatory tags should include paper number, title, authors and abstract text.
- Specify 8 1/2 by 11 inch page size for postscript papers.
- Urge submissions in a few selected word processor formats, provide style sheets to authors.
- Include postscript files on the CD-ROM for users to print or view.
- Keep a database of electronic submissions, their formats, the e-mail addresses of the authors, etc.
- Do your own backups! (Trust No One).
- Request ftp or media instead of e-mail submissions, or set up a mail daemon to handle submissions on a daily basis.
- Make sure the review staff has time to do the review.
- Make sure you can comply with licensee/customer review requirements or get a waiver. We had a signed agreement with Adobe to allow 14-day review, which could have been a disaster if they had enforced it.

NOTE: NCSA informed us in September that our use of the Mosaic binaries on the CD-ROM requires commercial licensing through Spyglass. Spyglass will license the software for $.50 per copy for under 5,000 copies, $.30 for over 5,000. The fee would be $850 for the 1700 IGARSS CD's that were pressed. We are still discussing the license arrangement with IEEE, who holds the copyright and owns the intellectual property.